Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 13-21 recibido 1-05-2009; aceptado 5-06-2009

Detección de Expresiones Temporales TimeML en Catalán mediante Roles Semánticos y Redes Semánticas *

TimeML Temporal Expressions Detection for Catalan Using Semantic Roles and Semantic Networks

Héctor Llorens, Borja Navarro, Estela Saquete
Grupo de Investigación en PLN y Sistemas de Información
Universidad de Alicante, España
{hllorens,borja,stela}@dlsi.ua.es

Resumen: Actualmente, la representación y procesamiento computacional de la información temporal en las lenguas naturales está siendo objeto de gran interés para la comunidad científica. El principal esquema de anotación para representar la información temporal es el TimeML, que ha sido tomado como estándar por un gran número de investigadores. Sin embargo, los recursos disponibles son muy limitados, sobre todo para lenguas diferentes del inglés. En este trabajo analizamos el uso de redes semánticas y roles semánticos desde una perspectiva multilingüe para la detección automática de expresiones temporales siguiendo el estándar TimeML. La propuesta ha sido evaluada para el catalán obteniendo un $F_{\beta=1}$ estricto de 83.7 %, y comparada con sus resultados para el inglés y el español confirmando que puede ser aplicada con éxito a diferentes idiomas.
Palabras clave: TimeML, roles semánticos, Catalan WordNet, AnCora

Abstract: Nowadays, representation and processing of temporal information in natural language is receiving a great research interest. The main annotation scheme for representing temporal information is TimeML, which has been adopted as standard by a large number of researchers. However, available TimeML resources are very limited, specially in languages other than English. In this work we analyze the usage of semantic networks and semantic roles from a multilingual perspective for the automatic detection of temporal expressions following TimeML specifications. This approach has been evaluated for Catalan obtaining an 83.7 % strict $F_{\beta=1}$, and compared to its results for English and Spanish confirming that it can be successfully applied in different languages.
Keywords: TimeML, semantic roles, Catalan WordNet, AnCora

1. Introducción

La importancia de los aspectos temporales del lenguaje natural (LN) no es algo nuevo en el campo de la inteligencia artificial (Allen, 1983). En los últimos años, la investigación en el tratamiento automático de la información temporal en el LN ha experimentado un gran crecimiento (Schilder, Katz, y Pustejovsky, 2007). Una de las principales razones para ello es la utilidad de esta información en diferentes áreas del procesamiento del lenguaje natural (PLN), tales como la búsqueda de respuestas y el resumen automático. La importancia de este campo se ve reflejada en numerosos talleres y conferencias especializadas (Pustejovsky, 2002), así como en foros de evaluación (TERN, 2004; Verhagen et al., 2007).

Hay diferentes maneras de representar la información temporal en LN. Entre ellas destaca TimeML (Pustejovsky et al., 2003a), adoptado recientemente como estándar para anotar eventos y expresiones temporales (ETs) por un gran número de investigadores.

Otra manera diferente de representar la información temporal es a través de los roles semánticos (Gildea y Jurafsky, 2002). Éstos la representan típicamente mediante el rol temporal.

Por otro lado, el tratamiento de información multilingüe se ha convertido en un tema destacado en la comunidad del PLN. Prue-

* Este artículo ha sido financiado por el Gobierno Español: proyecto TEXT-MESS (TIN-2006-15265-C06-01) donde H.Llorens está becado (BES-2007-16256)

ba de ello son conferencias como el CLEF[1], y trabajos en el reconocimiento multilingüe de ET como Wilson (2001) o Moia (2001).

Actualmente, el mayor problema del TimeML es la escasez de corpus para diferentes lenguas. La motivación de este trabajo reside en el estudio de mecanismos para ayudar en la creación de recursos TimeML a través de la explotación de los recursos multilingües disponibles, tales como los corpus anotados con roles semánticos y las redes semánticas. Para ello, se presenta un método automático que identifica ETs utilizando estos recursos desde una perspectiva multilingüe. Concretamente, en este trabajo se analiza la aplicación de este método al catalán. Finalmente, se presentan los resultados para dicho idioma y se comparan con los del inglés y el español.

El trabajo se estructura como sigue. La sección 2 analiza el estado de la cuestión en cuanto a TimeML y la aplicación de roles semánticos y redes semánticas al reconocimiento de ETs. En la sección 3 se detalla el método automático para obtener ETs TimeML mediante roles semánticos y redes semánticas. La sección 4 incluye la evaluación y análisis de errores del mismo para el catalán, así como una comparativa con otros idiomas. Finalmente, se presentan las conclusiones y trabajos futuros.

2. Estado de la cuestión

TimeML (Pustejovsky et al., 2003a) es un lenguaje de especificación para eventos, ETs y sus relaciones en LN. Combina y extiende características de los estándares de anotación anteriores, STAG (TIMEX) (Setzer y Gaizauskas, 2000) y TIDES (TIMEX2) (Ferro et al., 2005), lo que lo convierte en un esquema de anotación más potente. Los elementos básicos del TimeML son:

- **TIMEX3**, que marca expresiones temporales ("*mayo, lunes, dos años, etc.*").
- **EVENT**, que marca eventos ("*escapó, ataque, etc.*").
- **SIGNAL**, que marca señales temporales ("*en, durante, antes de, etc.*").
- **TLINK**, **ALINK** y **SLINK**, que marcan relaciones temporales, de subordinación y aspectuales entre los elementos anteriores.

La Figura 1 ilustra un ejemplo de anotación en TimeML. En el ejemplo, "*vino*" representa un evento que está relacionado a la expresión temporal "*junio*" a través de un enlace temporal (TLINK), en el que está involucrada la señal temporal "*en*".

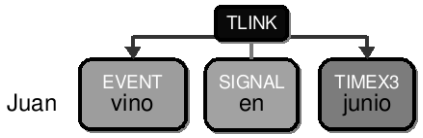

Figura 1: Ejemplo de TimeML

Siguiendo el esquema TimeML se creó para el inglés el corpus TimeBank (Pustejovsky et al., 2003b). La última versión, TimeBank 1.2, se considera actualmente como corpus de referencia y está publicada por el *Linguistic Data Consortium*. Desafortunadamente, no hay corpus TimeML para otros idiomas como el catalán.

Existen diferentes trabajos sobre sistemas de detección automática de TIMEX3. Por una parte, el sistema TTK (Verhagen et al., 2005) lleva a cabo esta tarea utilizando el módulo GUTime. No ha sido evaluado para TIMEX3 sino con la versión anterior, para la que obtuvo un 78 % en F para la identificación estricta en inglés. Por otra parte, Boguraev y Ando (2007) presentaron un sistema de reconocimiento de TIMEX3 para el inglés sobre el TimeBank utilizando técnicas de aprendizaje automático. Se obtuvo un 81.7 % en F para identificación estricta.

Como se introdujo en la sección anterior, otra forma de representar la información temporal en los textos en LN son los roles semánticos. Éstos determinan los eventos de una oración, detectando relaciones semánticas entre eventos y entidades. Todos los predicados semánticos se identifican y clasifican como argumento (agente, paciente, etc.) o adjunto (locativo, temporal, etc.), siendo principalmente el rol temporal el que representa la información temporal. La Figura 2 muestra un ejemplo de anotación con roles.

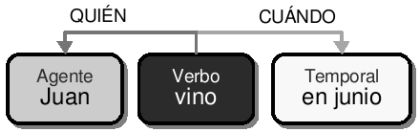

Figura 2: Ejemplo de roles semánticos

Sólo se ha encontrado una referencia sobre el uso de los roles semánticos para el procesa-

[1] http://www.clef-campaign.org/

miento de la información temporal: Hagège y Tannier (2007). En este trabajo se utilizan los roles como información complementaria para detectar relaciones temporales.

Por otro lado, las redes semánticas han sido ampliamente utilizadas en PLN con diferentes propósitos. WordNet (Fellbaum, 1998) es la red semántica más conocida y utilizada para el inglés, pero existen redes semánticas tipo WordNet para multitud de idiomas. Una lista de los WordNets disponibles en cada idioma puede ser consultada en la web de Global WordNet[2].

Para la tarea de identificación de expresiones temporales, las redes semánticas han sido utilizadas en los siguientes trabajos. En Negri y Marseglia (2004) se utilizó WordNet para crear una lista de entidades nombradas temporales tales como "*Bastille Day*", "*Hanukkah*", etc., extrayendo todos los hipónimos del *synset* "*calendar_day*". Por otro lado, en Saquete et al. (2004), las redes semánticas fueron utilizadas para expandir una lista de disparadores temporales añadiendo todos los sinónimos. Estos trabajos muestran que la información contenida en las redes semánticas puede ser útil para la extracción de información temporal.

3. Propuesta

En este trabajo se presenta un método automático que identifica expresiones temporales TimeML, desde una perspectiva multilingüe, utilizando roles semánticos y redes semánticas. Se implementan dos versiones del método: TIPSem que utiliza sólo roles semánticos y TIPSem+WN que los combina con diferentes WordNets.

3.1. TIPSem

El rol temporal no se corresponde exactamente con las expresiones temporales TIMEX3. Un rol semántico temporal (TSR) representa un predicado semántico completo con función temporal. Sin embargo, de acuerdo con las especificaciones del TimeML, la extensión de la etiqueta TIMEX3 debe corresponderse con una de las siguientes categorías: sintagma nominal (NP), sintagma adjetival (ADJP) o sintagma adverbial (ADVP). Como se muestra en el ejemplo 1, ambas representaciones no son equivalentes.

(1) Ell anà [en 1999 TSR][3]
Ell anà en <TIMEX3>1999</TIMEX3>

Para solucionar esto, ha sido creado un conjunto de reglas de transformación de TSR a TIMEX3. Originalmente se hizo para el inglés, pero con un enfoque multilingüe que permite ahora utilizarlo para otros idiomas como el catalán.

1. **Eliminación de la superposición**: Debido a que cada verbo de una oración tiene su propia anotación de roles, es posible encontrar TSRs superpuestos "*[quan eixires [ahir TSR] TSR]*"[4]. TIPSem sólo mantiene un TSR, el mínimo (NP, ADJP o ADVP).

2. **Eliminación de la subordinación**: Si un TSR corresponde a una subordinación, éste no representa un TIMEX3 "*quan vinga ell*"[5]. El sistema detecta la subordinación utilizando el árbol sintáctico y la elimina.

3. **Partición de TSR**: Un TSR compuesto por más de un NP puede contener varios TIMEX3 relacionados por preposiciones temporales o coordinaciones. Estas expresiones se marcan con etiquetas TIMEX3 independientes partiendo el TSR.

   Hay dos excepciones para esta regla. En primer lugar, las horas como "*les quatre i mitja*"[6], donde la conjunción "*i*" forma parte de la expresión temporal. Y en segundo lugar, la preposición "*de*" ("*final de mes*"), que denota una especificación. TIPSem busca preposiciones y coordinaciones en cada TSR que contenga más de un NP. Si son encontradas y no representan una excepción, el TSR es dividido en tantos TIMEX3 como NPs contenga, excluyendo tanto las preposiciones como las conjunciones de coordinación.

4. **Reducción sintáctica de TSR**: Un TSR difiere de un TIMEX3 en su tamaño. Si un TSR no se corresponde a la unidad sintáctica mínima, debe ser reducido. Los casos más comunes, como el del ejemplo 1, son aquellos en que el

[2] http://www.globalwordnet.org/

[3] "*Él fue en 1999*"
[4] "*cuando saliste ayer*"
[5] "*cuando venga él*"
[6] "*las cuatro y media*"

TSR consiste en un sintagma preposicional (PP). El PP contiene una preposición ("*en*") o una combinación adverbio-preposición ("*abans de*"[7]) que se corresponde con una señal temporal, seguido por un NP, que representa el TIMEX3.

5. **Transformación de TSR resultantes a TIMEX3**: Finalmente, los TSR resultantes se etiquetan como TIMEX3.

Debido a que los roles semánticos se apoyan en los verbos, las oraciones nominales no pueden ser etiquetadas. Estas oraciones se encuentran normalmente en títulos, paréntesis, notas, etc. Por tanto, como un paso posterior, se ejecuta, un reconocedor de fechas y horas explicitas ("1999", "18:25", etc.).

3.2. TIPSem+WN

Existen casos en los que un TSR no contiene ningún TIMEX3. Estos casos representan uno de los principales problemas del sistema TIPSem. El ejemplo 2 ilustra el problema mostrando una oración anotada con TSR, la anotación correcta en TIMEX3 y la anotación incorrecta que produce TIPSem.

(2) **TSR**:
Ell menjà [abans del viatge TSR][8]
TIMEX3 correcto:
Ell menjà abans del viatge
TIMEX3 incorrecto (TIPSem):
Ell menjà abans del
<TIMEX3>viatge</TIMEX3>

Como se muestra en el ejemplo 2, "*viatge*" es incorrectamente anotado como TIMEX3. En este caso la información temporal que aporta el TSR se corresponde con un evento según el TimeML. La dificultad reside en cómo diferenciar este tipo de eventos de las expresiones temporales. El siguiente ejemplo ilustra porqué esta distinción no es trivial utilizando únicamente la información morfosintáctica y los roles semánticos.

(3) (S(NP (PRP Ell))
(VP (VBD menjà)
(PP#TSR (RB abans)(IN del)
(NP (NN migdia[9])))))

(S(NP (PRP Ell))
(VP (VBD menjà)
(PP#TSR (RB abans)(IN del)
(NP (NN viatge)))))

En el ejemplo 3, "*abans del migdia*" y "*abans del viaje*" son representados por un TSR a nivel de roles semánticos, y su análisis morfosintáctico es idéntico. Sin embargo, "*migdia*" es una expresión temporal y "*viatge*" no.

Una posible solución sería construir manualmente una lista de disparadores temporales. Sin embargo, esta solución es muy dependiente del idioma. Por esta razón, proponemos una solución automática al problema utilizando la información multilingüe contenida en las redes semánticas disponibles para los diferentes idiomas. En particular, en este trabajo para el catalán se ha utilizado Catalan WordNet (Benitez et al., 1998).

Para cada sentido de una palabra (*synset*), las redes semánticas proporcionan, entre otras cosas, la jerarquía de hiperónimos. Nuestra hipótesis es que todas las palabras relacionadas con la temporalidad incluirán algún concepto temporal general en su jerarquía de hiperónimos. El ejemplo 4 muestra dos palabras relacionadas con un concepto temporal general en Catalan WordNet.

(4) hora (jerarquía de hiperónimos)
=> **unitat_de_temps**[10]
=> quantitat[11]
=> abstracció[12]

dilluns[13] (jerarquía de hiperónimos)
=> dia_de_la_setmana[14]
=> dia_natural[15]
=> **període**[16]
=> quantitat
=> abstracció

La única excepción que se incluye en esta hipótesis son las expresiones temporales puramente numéricas tales como fechas y horas ("12-11-1999", "18:25", etc.).

Considerando la presente hipótesis, definimos el siguiente algoritmo de validación de

[7] "*antes de*"
[8] "*Él comió antes del viaje*"
[9] "*mediodía*"
[10] "*unidad de tiempo*"
[11] "*cantidad*"
[12] "*abstracción*"
[14] "*día de la semana*"
[15] "*día natural*"
[16] "*periodo*"

ETs basado en redes semánticas.

- Un TSR es validado como TIMEX3 si al menos una de sus palabras posee un hiperónimo que coincide con un concepto temporal general o se trata de una fecha/hora numérica.
- Para tratar palabras polisémicas, por una parte, se utiliza la categoría morfológica para consultar las redes semánticas, y por otra parte, en palabras polisémicas de la misma categoría, si al menos uno de los *synsets* esta relacionado con un concepto temporal, el sistema lo valida. La razón para esta segunda condición es que si uno de los *synsets* está relacionado con el tiempo y la palabra está contenida por un rol temporal, entonces el sentido temporal probablemente sea el correcto.
- El algoritmo considera expresiones multipalabra para validar expresiones temporales compuestas como "*Corpus Christi*" o "*Sant Josep*"[17].

En este trabajo se ha implementado el algoritmo para el catalán tomando como conceptos temporales generales: *període*, *unitat_de_temps* y *temps*[18].

La Figura 3 ilustra la arquitectura final del sistema TIPSem+WN.

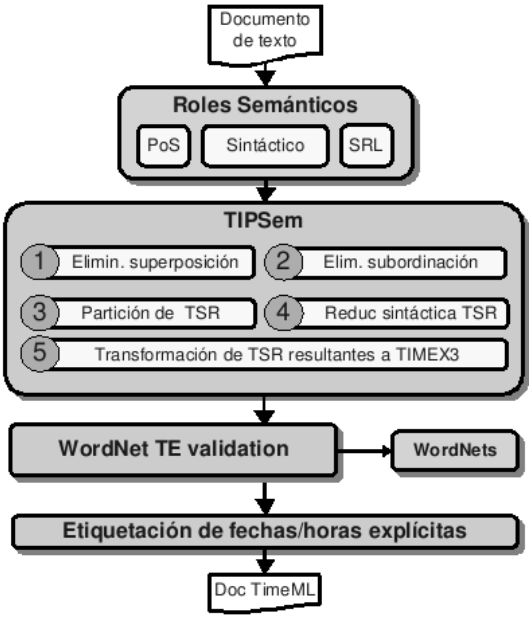

Figura 3: Arquitectura TIPSem+WN

4. Evaluación

El objetivo de la evaluación es estudiar la eficacia de los sistemas TIPSem y TIPSem+WN en la identificación de TIMEX3 en catalán. Para ello, se analizan los resultados obtenidos para este idioma comparándolos con los resultados obtenidos para el inglés y el español en trabajos previos (Llorens, Navarro, y Saquete, 2009). La evaluación incluye, además, un sistema *baseline* que etiqueta todos los TSR directamente como TIMEX3 para ampliar la comparativa.

4.1. Entorno de evaluación

4.1.1. Corpus

Debido a que no está disponible ningún corpus TimeML para el catalán, se ha desarrollado una muestra de corpus TimeML TIMEX3 anotando manualmente 30 documentos del corpus AnCora-Ca (Taulé, Martí, y Recasens, 2008; Martí et al., 2007). Esta muestra ha sido anotada sólo con el propósito de llevar a cabo la evaluación. AnCora es el corpus más grande accesible anotado a mano para el castellano (AnCora-Es) y el catalán (AnCora-Ca). Está compuesto por documentos de carácter periodístico anotados y revisados manualmente a nivel morfológico (PoS y lemas), sintáctico (constituyentes y funciones), y semántico (argumentos, roles, entidades nombradas y sentidos de WordNet).

El Cuadro 1 muestra las estadísticas de la muestra anotada. En el Cuadro, el campo *in TEXT* indica las etiquetas TIMEX3 en el cuerpo de los documentos (entre las etiquetas *TEXT*), ignorando las fechas explícitas de creación del documento, indicadas en las cabeceras.

documentos	**palabras**	**TIMEX3** (in TEXT)
30	8.0K	168 (138)

Cuadro 1: Estadísticas del corpus

4.1.2. Criterio

Los sistemas presentados se evalúan en la identificación de TIMEX3 en el corpus previamente descrito y los resultados se comparan con la anotación original. Las fechas explícitas de creación de los documentos se han ignorado para conseguir una evaluación más fiable. Para evaluar el rendimiento de los

[17] "*San José*"

[18] *periodo*, *unidad_de_tiempo* y *tiempo*

sistemas se ha aplicado el criterio utilizado en el TERN-2004. Las medidas tomadas del mismo son:

- **POS**: Total de etiquetas TIMEX3
- **ACT**[19]: Etiquetas TIMEX3 devueltas por el sistema.
- Correct (**corr**): Instancias correctas
- Incorrect (**inco**): Instancias correctas pero mal delimitadas
- Missing (**miss**): Instancias no detectadas
- Spurious (**spur**): Falsos positivos
- Precision (**prec**): corr/ACT
- Recall (**rec**): corr/POS
- $\mathbf{F}_{\beta=1}$: (2*prec*rec)/(prec+rec)

Para el cálculo de estas medidas se ha utilizado una adaptación a TIMEX3 del *scorer* del TERN-2004[20], originalmente desarrollado para el TIMEX2.

4.2. Resultados

Los Cuadros 2 y 3 muestran los resultados obtenidos para el catalán. Para cada sistema se indican los resultados en la identificación relajada $\boldsymbol{R}$ e identificación estricta $\boldsymbol{S}$. En R se consideran como correctas todas las etiquetas que identifican TIMEX3 aunque estén incorrectamente delimitadas, mientras que en S se requiere la coincidencia exacta de ambos límites de la expresión temporal.

System		corr	inco	miss	spur
Baseline	R	104	0	34	64
	S	65	39	34	64
TIPSem	R	116	0	22	44
	S	111	5	22	44
TIPSem+WN	R	113	0	25	7
	S	108	5	25	7

Cuadro 2: Resultados para el catalán *(1)*

System		prec	rec	$\mathbf{F}_{\beta=1}$
Baseline	R	61.9	75.4	**68.0**
	S	38.7	47.1	**42.5**
TIPSem	R	72.5	84.1	**77.9**
	S	69.4	80.4	**74.5**
TIPSem+WN	R	94.2	81.9	**87.6**
	S	90.0	78.3	**83.7**

Cuadro 3: Resultados para el catalán *(2)*

En la evaluación para el catalán, el *baseline* obtiene un 68 % de $\mathrm{F}_{\beta=1}$ para la identificación relajada pero baja hasta un 42.5 % en la estricta. El sistema TIPSem alcanza un 77.9 % y un 74.5 % de $\mathrm{F}_{\beta=1}$ para la identificación relajada y estricta respectivamente. Finalmente, el TIPSem+WN supera a ambos obteniendo un 87.6 % y un 83.7 % en $\mathrm{F}_{\beta=1}$ para la identificación relajada y la estricta.

Los resultados del *baseline* indican que aún tomando los roles temporales como TIMEX3, se obtiene un rendimiento bastante bueno en identificación relajada, sin embargo, sufren una fuerte caída en la identificación estricta. Por otro lado, el sistema TIPSem obtiene resultados mucho más altos, lo que indica que las reglas de transformación implementadas han solucionado varias diferencias entre el rol temporal y el TIMEX3. Finalmente, centrándonos en el sistema TIPSem+WN, podemos observar que su aplicación ha mejorado los resultados. Este hecho indica que el método definido para la validación de TIMEX3 basado en redes semánticas cumple su objetivo. La introducción de este método favorece la precisión y reduce muy poco la cobertura.

No se han localizado en la bibliografía resultados sobre la identificación de TIMEX3 para el catalán. Los resultados más cercanos a esta evaluación son los presentados en Boguraev y Ando (2007) para el inglés utilizando el TimeBank. La aproximación presentada para el catalán obtiene resultados similares a los del estado de la cuestión para el inglés.

Para ampliar el análisis de resultados, en la Figura 4, se muestra una comparativa de los resultados de $\mathrm{F}_{\beta=1}$ estricto obtenidos por las aproximaciones TIPSem y TIPSem+WN en español e inglés.

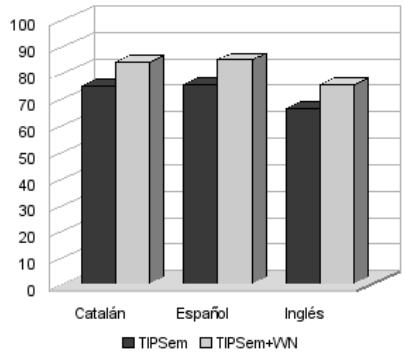

Figura 4: $\mathrm{F}_{\beta=1}$ estricto multilingüe

Como se puede observar, los resultados obtenidos para los tres idiomas evaluados siguen el mismo patrón y ofrecen una calidad

[19] Equivalente a Correct + Incorrect + Spurious
[20] http://fofoca.mitre.org/tern.html#scorer

similar. Las evaluaciones de español y catalán (AnCora) obtienen mejores resultados que la del inglés (TimeBank). Esto es debido a que, en el AnCora, la anotación de roles es manual, mientras que el TimeBank fue anotado con roles con una herramienta automática.

4.3. Análisis de errores

El objetivo de esta sección es mostrar en qué aspectos está fallando el TIPSem+WN y si los errores obtenidos en esta evaluación para el catalán son del mismo tipo que los obtenidos para el inglés y el español.

- **Falsos positivos (spurious)** (6 %): Este valor representa la cantidad de TIMEX3 etiquetados por nuestro sistema que no están etiquetados en el corpus de referencia. Este tipo de errores ha sido reducido drásticamente por la aplicación del método basado en redes semánticas. Concretamente, éste reduce los errores por falsos positivos del 27 % a 6 % en catalán. Esto confirma que la hipótesis planteada es válida para esta tarea en este idioma. Así mismo, para el inglés y el español esta reducción fue proporcional. Los falsos positivos restantes en las tres lenguas son expresiones temporales indefinidas, esto es, con un valor temporal indefinido ("un momento", etc.). El ejemplo 5 muestra algunas de ellas en catalán. El problema es que, aunque son expresiones temporales, no se corresponden con elementos TIMEX3 según las especificaciones TimeML.

  (5) l'últim moment[21]
  a temps[22]
  massa temps[23]
  temps de crisi[24]
  el moment de la jubilació[25]

- **Falsos negativos (missing)** (18 % CA): Este problema aparece porque los roles semánticos no siempre cubren todas las posibilidades de expresiones temporales en LN.

  - El mayor problema aparece en las oraciones nominales, paréntesis, y en general, todos los textos en los que los verbos no están presentes. Debido a que los roles semánticos principalmente están relacionados con los verbos, y a que el método basado en redes semánticas sólo se aplica a los roles temporales, el TIPSem+WN no es aplicable en estas oraciones. El ejemplo 8 ilustra estos errores.

    (6) una investigació de dos anys[26]
    (12 casos l'any passat)[27]
    endarreriments de mitja hora[28]
    les novetats d'aquest any[29]

  - Hay casos en los que una expresión temporal no tiene rol temporal en la oración pero es un TIMEX3. Principalmente son casos en los que la ET es argumento (agente, tema, etc.) o casos en los que la ET está a un nivel inferior al de argumento como complemento adverbial o del nombre. El ejemplo 7 ilustra este caso.

    (7) [Aquest any A0] [és V] [ara TSR], [amb les dades d'avui AM-MNR], [molt bo en comparació a l'any passat A1][30]

    En esta oración simple (con un solo verbo) queda reflejado cómo el rol temporal sólo se corresponde con la ET "*ara*". En cambio la ET "*aquest any*" tiene rol argumento, la ET "*avui*" es un complemento del nombre "*dades*" y la ET "*any passat*" es parte de un complemento adverbial.

    El ejemplo 8 ilustra errores de este tipo en catalán. Las redes semánticas no han sido aplicadas a roles que no sean temporales porque la ambigüedad introduciría ruido, por ejemplo en casos como "*Sant Josep, Victòria Abril*[31]".

    (8) [60 anys A0] no passen en va[32]
    he guanyat [5 anys de vida A1][33]

[21] "*el último momento*"
[22] "*a tiempo*"
[23] "*demasiado tiempo*"
[24] "*tiempo de crisis*"
[25] "*el momento de la jubilación*"
[26] "*una investigación de dos años*"
[27] "*(12 casos el año pasado)*"
[28] "*retrasos de media hora*"
[29] "*las novedades de este año*"
[30] "*Este año es ahora, con los datos de hoy, muy bueno en comparación con el año pasado*"
[31] "*Victoria Abril*"
[32] "*60 años no pasan en vano*"
[33] "*he ganado 5 años de vida*"

rep el Goya [per cinc decennis de cine AM-CAU][34]
la classe d'ahir[35]
en compració a l'any passat[36]

- En catalán, como en español, existen expresiones temporales en las que participan verbos (véase ejemplo 9). Estas expresiones adverbiales deícticas han supuesto falsos negativos en la evaluación porque la regla de eliminación de subordinación es sensible a los verbos.

  (9) [fa V] 17 anys[37]
  l'estiu que [ve V][38]

- Un problema que se esperaba encontrar al aplicar las redes semánticas era el aumento de los falsos negativos debido a la posible incompletitud de las mismas en cuanto a información temporal. Al contrario de lo que se pensaba, muy pocas expresiones temporales correctas obtenidas por TIPSem han producido falsos negativos en TIPSem+WN, lo que indica que, el Catalan WordNet es lo suficientemente completo en información temporal para satisfacer las necesidades de esta tarea. El ejemplo 10 muestra los únicos errores encontrados.

  (10) el lustre vinent[39]
  l'endemà[40]
  ple franquisme[41]

- **Incorrectos (incorrect)** (4 % CA): Los errores de delimitación son principalmente producidos por expresiones temporales con información adicional como "*ininterromputs*" en el ejemplo 11.

  (11) després de 16 anys ininterromputs[42]
  5 anys de victòries[43]

[34] "*recibe el Goya por cinco décadas de cine*"
[35] "*la clase de ayer*"
[36] "*en comparación con el año pasado*"
[37] "*hace 17 años*"
[38] "*el verano que viene*"
[39] "*el próximo lustro*"
[40] "*mañana*"
[41] "*pleno franquismo*"
[42] "*después de 16 años ininterrumpidos*"
[43] "*5 años de victorias*"

5. Conclusiones

Este trabajo estudia la aplicación los roles semánticos y Catalan WordNet a la identificación de expresiones temporales en catalán siguiendo las especificaciones del TimeML. Para ello, dos sistemas automáticos, planteados desde una perspectiva multilingüe, han sido desarrollados (1) TIPSem, que no utiliza las redes semánticas, y (2) TIPSem+WN que sí lo hace. Estos sistemas junto a un *baseline* han sido evaluados en la identificación de TIMEX3 el catalán, y comparados con resultados anteriores para el español y el inglés.

El TIPSem+WN ha obtenido para el catalán un $F_{\beta=1}$ relajado de 87.6 % y estricto de 83.7 %, lo cual representa una gran diferencia con el *baseline* y una importante mejora del sistema TIPSem.

Comparando esta evaluación con la del mismo sistema en español e inglés, se puede observar que los resultados siguen el mismo patrón y ofrecen una calidad similar. Por tanto, podemos confirmar que la aproximación presentada es válida para las tres lenguas. Debido a que la propuesta está basada en los roles semánticos y la información multilingüe de las redes semánticas, podría ser también válida para otras lenguas europeas.

No se han encontrado resultados comparables en la bibliografía. Sólo a modo indicativo se observa que son de la misma calidad que los obtenidos para el inglés por los sistemas del estado de la cuestión.

La escasez de corpus TimeML para idiomas diferentes del inglés impide el uso de técnicas de aprendizaje automático. Los resultados obtenidos en este trabajo para la detección de ETs, sin necesidad de disponer de corpus TimeML, nos llevan a plantear los siguientes trabajos futuros. En primer lugar, además de la utilización del análisis de errores para mejorar la propuesta presentada, se propone su extensión al resto de elementos TimeML (EVENT, SIGNAL, etc.), donde los roles semánticos y las redes semánticas pueden ser, también, de gran utilidad. En segundo lugar se plantea la evaluación del sistema en otros idiomas, especialmente aquellos que sean más diferentes como el Euskera, para confirmar si puede ser considerada multilingüe. Finalmente, como objetivo general, se propone el uso de la propuesta, una vez extendida, como ayuda en la creación semiautomática de corpus TimeML en diferentes lenguas.

Bibliografía

Allen, James F. 1983. Maintaining knowledge about temporal intervals. *Communications of ACM*, 26(11):832–843.

Benitez, Laura, Sergi Cervell, Gerard Escudero, Monica Lopez, German Rigau, y Mariona Taulé. 1998. Methods and Tools for Building the Catalan WordNet. En *ERLA Workshop on Language Resources for European Minority Languages, LREC*.

Boguraev, Branimir y Rie Kubota Ando. 2007. Effective Use of TimeBank for TimeML Analysis. En *Annotating, Extracting and Reasoning about Time and Events*, páginas 41–58. Springer.

Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. MIT.

Ferro, Lisa, Laurie Gerber, Inderjeet Mani, Beth Sundheim, y George Wilson. 2005. TIDES Standard for the Annotation of Temp. Expr. Informe técnico, MITRE.

Gildea, Daniel y Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3).

Hagège, Caroline y Xavier Tannier. 2007. XRCE-T: XIP temporal module for TempEval campaign. En *TempEval (SemEval)*, páginas 492–495. ACL.

Llorens, Hector, Borja Navarro, y Estela Saquete. 2009. Using Semantic Networks to Identify Temporal Expressions from Semantic Roles. En *RANLP (Accepted)*.

Martí, M. Antonia, Mariona Taulé, Lluís Márquez, y Manuel Bertran. 2007. Anotación semiautomática con papeles temáticos de los corpus CESS-ECE. En *Procesamiento del Lenguaje Natural*, volumen 38.

Moia, Telmo. 2001. Telling apart temporal locating adverbials and time-denoting expressions. En *Proceedings of the workshop on Temporal and Spatial information processing*, páginas 1–8, NJ, USA. ACL.

Negri, M. y L. Marseglia. 2004. Recognition and Normalization of Time Expressions: ITC-irst at TERN 2004. Informe técnico, Information Society Technologies.

Pustejovsky, James. 2002. TERQAS: Time and Event Recognition for Question Answering Systems. En *ARDA Workshop*.

Pustejovsky, James, José M. Castaño, Robert Ingria, Roser Saurí, Robert J. Gaizauskas, Andrea Setzer, y Graham Katz. 2003a. TimeML: Robust Specification of Event and Temporal Expressions in Text. En *IWCS-5*.

Pustejovsky, James, Patrik Hanks, Roser Saurí, Anderw See, Robert J. Gaizauskas, Andrea Setzer, Dragomir R. Radev, Beth Sundheim, David Day, Lisa Ferro, y Marcia Lazo. 2003b. The TIMEBANK Corpus. En *Corpus Linguistics*.

Saquete, Estela, Patricio Martínez-Barco, y Rafael Muñoz. 2004. Automatic Multilinguality for Time Expression Resolution. En *MICAI*, volumen 2972 de *LNCS*.

Schilder, Frank, Graham Katz, y James Pustejovsky. 2007. *Annotating, Extracting and Reasoning About Time and Events*, volumen 4795 de *LNCS*. Springer.

Setzer, Andrea y Robert Gaizauskas. 2000. Annotating Events and Temporal Information in Newswire Texts. En *LREC 2000*, páginas 1287–1294, Athens.

Taulé, Mariona, M. Antonia Martí, y Marta Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. En ELRA, editor, *LREC*.

TERN. 2004. Time Expression Recognition and Normalization Evaluation Workshop (http://fofoca.mitre.org/tern.html).

Verhagen, Marc, Robert J. Gaizauskas, Mark Hepple, Frank Schilder, Graham Katz, y James Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. En *Proceedings of the 4th International Workshop on Semantic Evaluations*, páginas 75–80, Prague. ACL.

Verhagen, Marc, Inderjeet Mani, Roser Saurí, Robert Knippen, Seok Bae Jang, Jessica Littman, Anna Rumshisky, John Phillips, y James Pustejovsky. 2005. Automating temporal annotation with TARSQI. En *ACL*, páginas 81–84, NJ, USA. ACL.

Wilson, George, Inderjeet Mani, Beth Sundheim, y Lisa Ferro. 2001. A multilingual approach to annotating and extracting temporal information. En *Workshop on Temporal and Spatial information processing*. ACL.